MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using Mapability Correction and Multiscale Signal Processing Framework

(Alternative: MUSEQ: Enrichment Analysis from ChIP-Seq Experiments using Mapability Correction and Multiscale Signal Processing Framework?)

[CITATIONS ARE NOT ADDED, YET]

ABSTRACT:

SHORTEN

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has become the mainstream experimental methodology for probing genome-wide enrichment of DNA binding proteins and post translational modification of histones, or histone modifications (HMs). Following a ChIP-Seq experiment, the generated reads are mapped and it is necessary to computationally identify the enrichments in the read depth signal. Unlike the transcription factors that show punctate signal enrichments, the enrichments for many ChIP-Seq assays manifest at much larger length scales (e.g., histone modifications and Polymerase II binding) and the spectrum of length scales can vary significantly between different types of ChIP-Seq experiments. This, combined with the systematic noise added to the signal by the non-uniform mapability makes it challenging to process the enrichments in the HM signal profiles.

MORE FOR BROAD

In this paper, we present MUSIC, a novel method for identifying and processing enrichments in the genome-wide read depth signal profiles from ChIP-seq experiments. MUSIC uses a mapability correction procedure coupled with multiscale signal processing based approach to identify the significant enrichment features (SEFs) that represent the significant enrichments at different length scales in the signal. By piling up the SEFs, MUSIC generates a genome-wide signal that can be utilized for quantifying the broadness of enrichment at each location. We show the utility of SEFs and SEF pileup signal within two applications: First we identify the enriched regions (ERs) using the SEFs and compare the accuracy to other ER identification methods, we show that MUSIC performs favorably in terms of accuracy of ERs and reproducibility of the ERs to a ground truth compared to other methods.

MORE LOGIC

WHY

Second, in order to showcase a novel application of the SEF pileup signal, we concentrate on processing the ChIP-Seq data for Polymerase II binding. The broadness of enrichments in signal profiles for polymerase binding can be used as an indicator of the polymerase state, i.e., stalled or elongating. We process the SEF pileup signal at protein coding gene promoters and demonstrate the bimodality of the joint distribution of signal broadness at the promoter versus the gene expression levels for the protein coding genes. This observation points to the fact that the SEF pileup signal can be utilized for analysis of broadness of genomic signals that can reveal important biological information.

MUSIC is available for download from Gerstein Lab Github page at: https://github.com/gersteinlab/MUSIC

INTRODUCTION:

With the recent advancements in sequencing technologies, chromatin immuniprecipitation based enrichment of the DNA sequences followed by sequencing (ChIP-seq) has become the mainstream experimental method for genome-wide measurement of DNA binding proteins (e.g. transcription factors) and posttranslational modifications of histone proteins, or histone modifications (HMs). Following the sequencing, it is necessary to computationally process the read depth signal profile to analyze the enrichments. There has been a lot of work on identification of transcription factor binding sites, or peaks, from ChIP-Seq experiments (cite{Rozowsky,Kharchenko, Fejes,Jothi, MACS reference}). Identification of the broad enrichments in the read depth signal profiles, however, did not receive the same amount of attention. The broad enrichments are observed for most of the HMs signal profiles. HMs like H3k9me3 show enrichments that can extend upto megabases. Another interesting example is RNA polymerase II, which binds to the promoters and gene bodies for the purpose of mRNA transcription, can extend over the whole gene bodies. Identification and characterization of the broad enrichments is the basic first step for understanding the regulatory effects of the HMs and diffuse DNA binding proteins on gene expression as more evidence is brought to light that these epigenetic factors are major driving factors for disease manifestation like cancer.

NO

HARDER

Ozgun 1/16/14 3:41 PM
Comment [1]: Citations on epigenetics and disease

Several popular methods for identification of broad enrichments include change point identification within the formality of Bayesian inference (BCP), local island identification and clustering (SICER), local thresholding and merging (MACS), and using local Poisson statistics to identify broad enrichments (SPP), Wavelet based smoothing and identification of enriched regions (WaveSeq, Kharpikov et al). In the ENCODE project, the main concentration has been on building integrative unsupervised segmentation models \cite{SegWay\cite{XXX}, ChromHMM\cite{XXX} } for using these modifications to annotate and characterize the cis-regulatory elements in the genome and studying the relation between the HM levels at these elements and the gene expression. Although the segmentation methods proved very useful for identifying novel regulatory elements like enhancers, the identification of the enriched regions per HM signal profile and the technical issues like mapability have not been adequately addressed. Finally, some of the HMs are almost never processed in the segmentation methods process simply because they are not well characterized in terms of their function and in terms of how they interact with other HMs.

NO

There are two main challenges for identification of enrichments in the HM signal profiles. First is that unlike transcription factor binding, HMs enrichments are observed at much larger length scales and the spectrum of enrichment lengths are broad for different types of ChIP-Seq experiments. This makes it necessary to identify the enrichments at different scales. A naïve and widely used method for identifying the HM signal profiles is smoothing the signal profile with a kernel of constant size and shape and using a null model (e.g., Poisson or negative binomial based) to identify the significantly enriched regions. It is, however, not obvious how the kernel size and shape should be selected. The multiscale approaches proposed by the wavelet based methods address this aspect but in those approaches, the selection of the predefined wavelet functions are not justified for their choice. Second, the signal profiles contain systematic noise introduced to the read depth signal by the repeat regions with low mapability. This noise causes discontinuities in the identified enrichments. This becomes an important factor especially in the intergenic regions where there is non-uniform mapability.

Ozgun 1/16/14 12:38 PM
Comment [2]: Wavelet based methods are different from ours as they do not utilize the the natural merging of clusters generated by multiscale smoothing approaches

In this paper, we present MUSIC, a method to identify enriched regions (ERs) in ChIP-Seq experiments. We utilize a multiscale signal processing approach based on a novel multiscale median filtering decomposition to identify the regions of enrichment in the signal profile generated. MUSIC uses an

BIG FAT

efficient mapability correction at the nucleotide resolution before multiscale decomposition of the signal so as to correct for the spurious loss of signal because of low mapability. Unlike the wavelet based multiscale approaches that are based on linear filtering, we take an approach to multiscale decomposition using the non-linear edge preserving median filtering, which has not been applied to processing ChIP-Seq datasets before. Basically, MUSIC exploits the fact that the multiscale smoothing procedure reveals the enriched features in the signal as "blobs" that can be detected as the regions between consecutive local minima of the smoothed signal. These enriched features are then trimmed and filtered with respect to their significance. This procedure yields the significant enrichment features (SEFs), the output of MUSIC algorithm. By pileup of the SEFs, MUSIC generates a genome-wide signal, which quantifies the broadness of enrichment at each position. This enables us to study the broadness of signal enrichments from different ChIP-Seq experiment. We show additional utility of SEFs with two applications. First, to evaluate the accuracy of SEFs for identification of enriched regions (ERs), we concentrate on H3k36me3, a well characterized HM that gets enriched on expressed gene bodies, which we use as ground truth. We build ERs from SEFs and compare the accuracy of ERs with 5 other methods with respect to accuracy, in terms of consistency with expressed regions, and reproducibility. We show that ERs identified by MUSIC have higher F-measure and higher reproducibility compared to other methods.

LOGIC

POL CPLX

Next, we demonstrate a novel utility of SEF pileup signal for analysis of enrichment of RNA polymerase II binding. The broadness of the enrichment of Polymerase ChIP-Seq signal profiles close to gene promoters can be used to distinguish the stalled and elongating polymerase. The stalled polymerase shows a punctate binding, whereas the elongating polymerase shows a much more broad binding. In addition, the stalled polymerase does not transcribe the gene, that can be observed as the low or no detectable expression of the downstream gene. For assessing the broadness of binding of polymerase, we generated the SEF pileup signal for Polymerase II using the ChIP-Seq data from ENCODE project. We demonstrate that joint distribution of broadness of polymerase enrichment at promoter, as quantified by SEF pileup signal, versus gene expression shows a bimodal characteristic, where one mode corresponds to the stalled polymerases and other mode corresponds to the elongating polymerases. This showcases a novel application of SEFs and a novel benefit of MUSIC.

Remaining of the paper is as follows. We first describe the MUSIC algorithm and identification of SEFs. Next we present identification of ERs and compare the accuracy and reproducibility of MUSIC other ER identification methods. Then we focus on joint processing the signal profiles for Polymerase II and gene expression levels. Finally, we present the algorithmic details of MUSIC in Methods.

RESULTS:

We first present motivation for MUSIC algorithm and lay out the steps of the algorithm. Then we present comparison of MUSIC with other ER identification algorithms. We finally present a novel application of the enrichment features identified by MUSIC to the joint processing of the polymerase data with gene expression levels.

MUSIC ALGORITHM:

There are two factors that motivate the novel methodology behind MUSIC:

1. Mapability is an important aspect of read mapping and processing. For example, in the repetitive regions the number of uniquely mapable positions decreases significantly. This, depending on the mapping tool, causes a systematic decrease of signal at repetitive regions and makes it impossible to evaluate whether a decrease in the signal is due to low mapability or a decrease in the modification levels. This becomes problematic especially in the intergenic and intronic regions which contain many repetitive regions.

BROAD PUNC

In order to characterize the mapability of different regions, MUSIC generates the genome-wide multi-mapability signal profile. This signal profile is similar to other mapability maps computed previously. For each position, this profile contains the number of reads (of certain length) that can map from any other position in the genome. In order to gain a perspective on the statistics of multi-mapability signal, we aggregated the signal over different elements. This reveals, as expected that, that the protein coding exons and promoter regions show the highest mapability (smallest multi-mapability signal) with (Figure SXX), with average of 1.2 reads mapping at these positions on average. We computed multi-mapability profiles for several read lengths and they are available for download with MUSIC (See Methods).

It is worth noting that although there are computational methods that aim at assigning the reads uniquely to repeat regions by resolving the multi-mapping reads, the underlying algorithms are too compute intensive for the purpose of enrichment identification, where the computation power should be allocated for identification of the enriched features.

Ozgun 1/16/14 4:40 PM
Comment [3]: cite

BUT DO WE NEED TO BENCH AG.?

2. The length distribution of ERs for broad enrichments usually have a large variance. This makes it necessary to identify the enrichments for a spectrum of scales. For example for HMs like H3k36me3, H3k27me3, the ERs can extend from couple of kilobases to hundreds of kilobases. On the other hand, for HMs like H3k4me3 and H3k27ac, which marks the gene promoters and enhancers, the ERs are around couple kilobases in length. Another interesting example is the RNA Polymerase II, whose enrichments can extend from less than a kilobase to hundreds of kilobases.

Motivated by these facts, we designed MUSIC to account for the effects of mapability and to be scale sensitive. In essence, MUSIC first corrects the signal profile from ChIP experiment for the mapability then computes the multiscale decomposition of the signal by smoothing signal with at multiple scale levels, then uses the decomposition to compute the significant enrichment features. In the process of smoothing, fine details in the signal are removed and broad enrichments are revealed as "blobs" in the smoothed signal, which can be detected as the regions between consecutive local minima of the smoothed signal. With multiple scales, MUSIC can detect enriched features within a spectrum of lengths that can be tuned by adjusting the starting and ending scale levels to be processed by MUSIC.

[[Each scale level is represented with smoothing window length where higher scales are associated with longer smoothing windows and thus are associated with broader features.]]

WHY NOT WAVELET

Figure XX shows the flowchart of MUSIC. The details of each step can be found in Methods. Here we summarize each step briefly. The input to MUSIC are the sets of reads from the ChIP and control samples (Steps 1 and 2), and the set of smoothing window lengths to be used in multiscale analysis. MUSIC first preprocesses the reads and filters the PCR duplicates for both samples. Then MUSIC computes a scaling factor using linear regression between the ChIP and control signal profiles. The slope of the regression is used as a normalization factor for control.

Then, in Step 3, the ChIP and normalized control signal profiles are generated, and the ChIP profile is smoothed and corrected with respect to mapability using the multi-mapability profile. The correction can be formulated as following:

$$\tilde{x}_i = \overbrace{\max[x_i, \underbrace{\text{median}\big(\{x_a\}_{a\in[i-l/2,i+l/2]} \mid m_a < \overline{m}_{\text{exonic}}\big)}_{\substack{\text{Median of the signal values at highly mapable} \\ \text{positions around } i}}]}^{\substack{\text{Compare the signal value at } i \text{ with} \\ \text{the median signal at highly mapable positions}}}$$

Where x_i and $\tilde{x}_i$ are the uncorrected and corrected signal values, respectively, at position i, m_a is the value of multi-mapability profile at position a, l is the length of median filter utilized in correction which is set to 2000 base pairs, and $\overline{m}_{\text{exonic}}$ is the average multi-mapability value over the exonic regions, which we identitied as the most mapable regions in the genome. In summary, MUSIC first generates the mapability corrected signal profile from the ChIP-seq signal (See Methods), where for each position i, MUSIC first computes the median of the signal values at highly mapable positions (multi-mapability signal smaller than 1.2) within 2000 bp vicinity of i. Then MUSIC compares this value with the signal value at i and assigns the maximum of them to the corrected value. The basic idea behind this correction is that since we know that mapability causes loss of signal, if the signal value at i is higher than its vicinity, then it is highly likely that the mapability did not affect the signal value at i. It should be noted that maximum filtering, also known as dilation in image processing, is used for feature enhancement in images.

MUSIC then performs median filtering to the mapability corrected ChIP profile to compute multiscale decomposition of ChIP signal at multiple scales (Step 4.) For a given scale s with smoothing window length l_s, the median smoothed signal is formulated as:

$$x_i^s = \text{median}\left(\{\tilde{x}_a\}_{a\in\left[i-\frac{l_s}{2},i+\frac{l_s}{2}\right]}\right)$$

where x_i^s is the filtered value at position i. The median filtering is utilized extensively in signal processing as an edge preserving smoothing filter. MUSIC utilizes median filtering in a novel application in multiscale decomposition.

For each decomposition, MUSIC identifies all the local extrema, i.e., local minima and local maxima (Step 4). The regions between the consecutive local minima are marked as the candidate enrichment features.

It is worth noting that these features have exactly one local maxima in them. For each enrichment feature, MUSIC computes the fraction of the maximum of smoothed ChIP signal (at the corresponding scale) to the unsmoothed ChIP signal within the boundaries of the enrichment feature. If this fraction is smaller than the maximum signal smoothing threshold (denoted by γ), MUSIC discards this enrichment feature (Refer to Methods.) This way MUSIC removes, at large scales, the features with local enrichment.

The features identified from the consecutive minima are rough and it is necessary to identify the location of densest signal enrichment within each feature. To achieve this, MUSIC performs a Poisson background based thresholding and p-value minimization to trim the ends and identifies the densest regions of signal enrichment in the enrichment feature (Step 5 in Fig XX). For a feature starting at position i and ending at position j the trimming operation identifies new start and end coordinates, i' and j', can be formulated as following:

$$(i', j') = \underset{\substack{a,b,\\ i<a<b<j}}{\operatorname{argmin}}\left(p(a,b) \mid \exists a' \in (a,b) \text{ s.t. } x_{a'} > \tau\right)$$

where τ is the threshold identified from Poisson model, and $p(a,b)$ represents the binomial p-value of the region within coordinates a and b, and i' and j' represent the start and end of the trimmed feature, respectively. (Refer to Methods). Finally, MUSIC computes the binomial p-value for each trimmed enrichment feature and filters out those whose p-values are larger than 0.05. We refer to the remaining enrichment features as significant enrichment features (SEFs).

The basic assumption is that SEFs contain all information about the local enrichments in the signal over a spectrum of smoothing scales, therefore MUSIC utilizes the SEFs for processing the enrichments in the signal. Next, we will present the utility of SEFs with two applications.

2.

Evaluation of Enrichment Broadness with SEF Pileup Signal

The scale dependence of SEFs is a useful property for evaluating the “broadness” of enrichments. Each SEF represents a locally enriched region at a certain smoothing window length. Therefore, a position that is covered by many SEFs has a broader enrichment than a position that is covered by small number of SEFs. Following this basic observation, MUSIC generates the SEF pileup signal profile, a genome-wide profile that is generated by counting the number of SEFs covering each position, which also quantifies the broadness of enrichment at each position in the genome.

To illustrate this, we processed multiple ChIP-Seq datasets (CTCF, Polymerase 2, and several HMs) from ENCODE project for K562 cell line from with MUSIC for smoothing scales starting from 100 bp scale to 2.5 megabase scale with exponentially increasing smoothing window lengths of 1.5 (Total of 25 scales) and generated the SEF pileup signal for chromosome 1. Figure 2 shows the distribution of SEF pileup signal for different datasets. In this plot, we mapped the value of SEF pileup signal to its corresponding smoothing window length which is also shown in the x-axis. As expected CTCF, a punctate binding transcription factor, shows the least broad enrichments compared to other datasets. H3k4me3 and

H3k4me1, active promoter and enhancer HM marks, show broader enrichments than CTCF. H3k36me3 and H3k27me3, which mark active and repressed gene bodies, show broader enrichments and finally H3k9me3, an HM associated with large heterochromatin domains, shows the broadest enrichments. Another interesting observation is that H3k4me3, H3k4me1, and H3k36me3 have maxima at certain scales, which indicates that these HMs get enriched at specific length scales that are observed very frequently. Finally RNA Polymerase II signal profiles show a high frequency of enrichments at small scales and relatively high frequency of enrichments at large scales.

Identification of ERs Using SEFs

MUSIC utilizes SEFs to identify enriched regions in the genome. The ERs, unlike SEFs, is a set of non-overlapping regions that are enriched. For this, the candidate ERs are computed by merging the union of the SEFs identified from all the decompositions (Step 6 in Fig. 1). MUSIC then evaluates the quality of the ERs with respect to concordance of the signal levels on positive and negative strands. MUSIC computes the amount of signal mapping to positive and negative strand in each ER and filters out the ERs for which the fraction of total signal on positive strand to that on negative strand (or vice versa) is less than 0.5. We observed that this filter removes many spurious ERs for the HMs with relatively less broad enrichments (See Methods).

For each of the remaining ERs, MUSIC computes a binomial p-value using the number of reads in the ChIP and control samples. Since the features have different lengths, the all the counts are normalized to $l_{p_{val}}$ window length (See Methods):

$$p(i,j) = Bin(\frac{n_{chip}}{(j-i+1)} \times l_{p_{val}}, \frac{n_{control}}{(j-i+1)} \times l_{p_{val}})$$

Where n_{chip} and $n_{control}$ are the read counts in ChIP and control samples within the ER starting at nucleotide position i and ending at j. The p-values are corrected by Benjamini-Hochberg procedure\cite{XXX}. The final corrected p-values are thresholded with respect to 0.05 for identification of significant ERs. MUSIC can be utilized to determine ERs from precomputed SEFs ("-get_ERs_per_SEFs"), or identify ERs from reads ("-get_peaks_per_reads").

Comparison with Other ER Identification Methods:

In order to evaluate the accuracy of the enriched, we compared the ERs from MUSIC with 5 other algorithms that identify ERs from ChIP-Seq data: BCP, SPP, MACS, SICER, and PeakRanger. We ran all the algorithms using H3k36me3, and H3k27me3 ChIP-Seq datasets for GM12878 and K562 cell lines from ENCODE project. H3k36me3 correlates well with expressed transcript regions and this allows us to build a ground truth set for H3k36me3 as the bodies of expressed transcripts. We downloaded the transcript quantifications (in RPKMs) from ENCODE RNA-seq dashboard and thresholded the expression levels of the transcripts and filtered the transcripts with low expression. The expressed transcripts are then merged to generate the final set of expressed regions. Rather than selecting one expression threshold,

Ozgun 1/15/14 6:11 PM
Comment [4]: Add citations to this

we selected thresholds between 0 and 1 increasing with steps of 0.01 so as to evaluate the accuracy of peak calls against multiple ground truth sets identified at different levels of expression.

Accuracy Measures:

To measure the accuracy of ERs, we used sensitivity (the fraction of the coverage of correctly predicted ERs to the coverage of the ground truth set) and positive predictive value (the fraction of the coverage of correctly predicted ERs to the coverage of identified ERs). In order to combine the sensitivity and PPV into one accuracy measurement, we used F-measure, which is the harmonic mean of sensitivity and positive predictive value (See Methods). Having one measure of accuracy enables us to easily compare the accuracy of methods with changing RPKM thresholds.

Figure 3a and b shows the F-measure of the H3k36me3 peak calls for different methods with respect to the changing RPKM cutoffs used to identify expressed regions. MUSIC has higher F-measure than all the other methods for GM12878 at all expression cutoffs, followed by BCP. For K562, MUSIC has higher F-measure than all other methods for expression cutoffs smaller than 0.8 then falls slightly below BCP. It should be noted that RPKM cutoff of 0.8 is a very stringent threshold for identifying expressed transcripts.

For assessing the importance of mapability correction, we ran ER identification without mapability correction and computed the F-measure of the ERs. Fig 3c shows the F-measure versus RPKM threshold. Using mapability map significantly increases the accuracy of peak calls and shows the importance of utilizing the mapability correction in ER identification.

[[TIME AND MEMORY USAGE COMPARISONS]]

Analysis of Polymerase II Enrichments and Gene Expression Levels using SEFs

Next, in order to illustrate a novel utility for the SEFs identified by MUSIC, we concentrated on the Polymerase II binding data from ENCODE project. Polymerase shows distinct patterns of binding such that the depending on the state of polymerase, i.e., elongating or stalled, the ChIP-Seq enrichment becomes more broad and more punctate for elongating and stalled polymerase, respectively. In addition, the stalled and elongating polymerase can be distinguished by comparing the detected amount of transcription at the polymerase binding.

For evaluating the relation between the expression and the enrichment broadness as measured by SEF pileup signal, we processed and computed the SEF pileup signal (100 bases to 2.5 megabases) using the ChIP-Seq dataset for RNA polymerase II (Pol2b) from ENCODE project. For each protein coding gene, we computed the maximum value of the SEF pileup signal within the promoters. This gives us, at each gene, an estimate of the broadness of polymerase binding at the promoter. Next, we also quantified the gene expression levels in RPKMs. Finally, we plotted the joint distribution of SEF signal and gene expression level for each gene which is plotted in Fig. 5. Visual inspection of this plot reveals two components: The maximum of one component can be located at SEF pileup signal at 9 and log expression (log expression level at 2. This component can be associated with actively transcribed genes. The maximum of other component is located at SEF pileup signal at 9 and log expression level at 0. Although the maximum does not have a distinguishable local maximum, it can be spotted by looking at the distribution from two different orientations, as in Fig. 5a and 5b.

[[FOLLOWING IS THE AGGREGATION PLOT OF Pol2s2 data: 4 quandrants in the expression/broadness plane]]

DISCUSSION:

We present a novel method, MUSIC, for identification of enriched regions in ChIP-Seq experiments. Although MUSIC can be used to identify enrichments in any ChIP-Seq experiment, we concentrated on identifying histone ChIP-Seq experiments in this paper. MUSIC utilizes a multiscale decomposition of the ChIP-seq signal profile in conjunction with a novel mapability correction filtering for the correction of the data. Mapability is an important aspect of peak calling from next generation sequencing data especially for identifying the broad domains of enrichment since the read depth profiles are highly correlated with the mapability map. MUSIC, to our knowledge, is the first peak caller that takes mapability into account for identifying broad domains of enrichment at nucleotide level. We showed that MUSIC outperforms other methods in terms of accuracy of H3k36me3 peaks in comparison with the expressed transcripts identified from the expression data from ENCODE project.

An important advantage of MUSIC is that the users can specify the peak calling scale that they would like to process, which can be easily done using the scale parameters for the multiscale decomposition, which sets the scale at which the algorithm identifies the enrichment features. To our knowledge, other peak calling methods do not present an intuitive way to set the scale at which the peaks are called. We believe this customizability will prove very useful for processing the datasets generated using ChIP-Seq experiments for which broad binding profiles are observed.

[[There is no mode for one sample analysis, which is reasonable bc …]]

As with all algorithms, MUSIC has limitations. Currently, MUSIC cannot be directly used on genomes with high chromosomal aberrations, i.e., copy number variations. Although the Poisson background model partly compensates for this by modeling the read distribution over a large window, the current significance estimation by binomial p-value computation does not correct for these effects and can therefore generate spuriously high number of peak calls on regions with high copy numbers. This is a limitation that is not addressed by many peak callers and is vital to perform epigenomics analyses on the genomes with extreme copy number variations like cancer samples.

We believe that MUSIC is an important tool for analysis for identification and analysis of enrichments in ChIP-Seq datasets.

[[MUSIC allows changing the smoothing scale levels used for analyses in an intuitive manner, which is a novel feature]]

METHODS:

We describe signal processing pipeline underlying MUSIC in more detail.

REF ?

Control Scaling Value Computation:

Mapability Correction and Enrichment Feature Enhancement: Given the read depth signal at each nucleotide position, MUSIC generates the per nucleotide multi-mapability signal and corrects for the mapability based loss of signal using following filtering:

$$\tilde{x}_i = \max\left[x_i, \mathrm{median}\left(\{x_a\}_{a\in[i-l/2,i+l/2]} \mid m_a < 1.2\right)\right]$$

Where x_i is the signal value at nucleotide position i, $median(\{x_i\})$ is the median of the set $\{x_i\}$, m_a is the value of the multi-mapability profile at the position a, and l is the window length used in mapability aware filtering. Using this filtering, MUSIC infers the signal values for positions with low mapability using the median of the values at nearby positions with multi-mapability signal lower than 1.2. We selected this value since it is the smallest multi-mapability signal profile value, i.e. most mapable, over exons and promoters as shown in Fig XXX. We set the window length l to 2000 bps from observations. This window length depends on the distribution of length of the non-mapable region lengths. Different window lengths did not seem to affect the results too much for our tests on human genome.

This filtering is inspired from the dilation operation in image processing, which is a morphological filter and has been used, in combination with other filters, for image enhancement. In our experiments, we observed that the operation defined above tends to enhance the significant features and does not change the significance of the background regions.

Multiscale Enrichment Feature Identification: Multiscale signal processing has been used in the context of wavelet transform~\cite{XX,XX,XX} to process ChIP-Seq data and for peak calling. In this paper, we are using a more general form of multiscale filtering, namely the multiscale decomposition~\cite{XX}. MUSIC utilizes a median filtering based smoothing for generating a multiscale decomposition. We selected to use median filtering since it has many applications in signal processing for performing signal smoothing with edge preserving. Given a window length, i.e. the scale, median filtering can be formulated as:

$$x_i^s = \mathrm{median}\left(\{\tilde{x}_a\}_{a\in\left[i-\frac{l_s}{2},i+\frac{l_s}{2}\right]}\right), l_s \in (l_{begin}, l_{end})$$

Where x_i^s is the i^{th} value of the decomposition at scale level s for which the smoothing window length is l_s, and $\tilde{x}$ is the mapability corrected signal profile. The window length l_s, that MUSIC uses has an exponential increase~\cite{XX} to make sure that the larger scales do not dominate the generated features.

The multiscale decomposition enables automatic identification of blobs in the signal profiles at different scales with very small computational requirement. MUSIC uses a fast and efficient method to implement the median filtering by storing the histogram of the signal values in the window and processes only the new and obsolete signal values that enter and leave, respectively, the current window to update the histogram when moved to the next window.

It should be noted that any type of smoothing filter can be used, e.g., Gaussian, Triangular, Rectangular, etc., to generate the multiscale decomposition for peak calling. MUSIC currently supports three filters: Median, Gaussian, and Mean. To our knowledge MUSIC is the first algorithm to utilize a non-linear median filtering for multiscale feature identification for processing genomic signal profiles.

[[ADD PER SCALE SIGNAL LOSS FILTER: Maximum signal smoothing threshold (denoted by γ)

]]

Identification of Enrichment Features: After the multiscale decomposition, MUSIC identifies all the local minima points in the decomposition. MUSIC utilizes regions between minima points as the regions of enrichment. We refer to these regions as *enrichment features*.

Feature End Trimming using Poisson Distribution Model: MUSIC trims the ends of the features first using a Poisson null model for the signal distribution. For this, MUSIC first divides genome into 1 megabase windows and for each 1 megabase window estimates the mean of all the values. Using this as the mean parameter μ of the Poisson distribution, MUSIC selects a threshold that satisfies 5% false positive rate:

$$\tau = \underset{t}{\operatorname{argmin}}\{F_{X_\mu}(t) > 0.95\}\ , X_\mu \sim Poisson(\mu))$$

Where F_{X_μ} represents the cumulative distribution function of X_μ , which is distributed as Poisson with mean μ. For a feature with start and end at positions i and j, respectively, the trimmed end coordinates are given as:

$$i' = \underset{a}{\operatorname{argmin}}(x_a > \tau), a \in (i,j)$$

$$j' = \underset{a}{\operatorname{argmax}}(x_a > \tau), a \in (i,j)$$

Where i' and j' are the trimmed start and end coordinates, respectively. The features that do not pass the threshold are removed from the candidate peak list.

Feature End trimming via p-value minimization: Then MUSIC further fine-tunes the ends of the merged features using a novel p-value minimization using the chip and control profiles. For a given region, MUSIC starts thresholding the signal at the ends of the region and identifies the signal height at which the p-value of the region is minimized. This maximizes the compactness of the merged feature regions.

The end-refined merged feature regions are the candidate regions of enrichment before p-value computation.

$$i' = \underset{a}{\operatorname{argmin}}\Big(p(a, j \mid l_{p_{val}} = (\mathrm{j} - \mathrm{a} + 1))\Big), a \in (i, j)$$

$$j' = \underset{a}{\operatorname{argmin}}\Big(p(i', a \mid l_{p_{val}} = (\mathrm{a} - \mathrm{i}' + 1))\Big), a \in (i', j)$$

where $p(a, b \mid l_{p_{val}})$ represents the p-value for the peak starting at a and ending at b with the length of p-value window given by $l_{p_{val}}$ (Refer to p-value computation.)

Feature Merging: After the features are identified, MUSIC merges all the features and identifies where the clumps of features are. This is done basically by identifying the positions that are covered by at least 1 feature.

Per Strand Concordance Filter: For each ER, MUSIC computes the total signal on positive and negative strands and filters the peaks for which there is high discordance between the signals:

$$\min\left(\frac{\sum_i x_i^+}{\sum_i x_i^-}, \frac{\sum_i x_i^-}{\sum_i x_i^+}\right) < 0.5$$

where $\sum_i x_i^+$ and $\sum_i x_i^-$ is the total signal on the positive and negative strand within the start and end coordinates of ER, respectively.

P-value and FDR Correction: We use one-tailed binomial test to compute the p-values for each end-refined merged feature region. We first count the number of reads in the chip sample (n_{chip}) and control sample ($n_{control}$) that overlap with the region, then compute one tailed p-value as:

$$p = \sum_{r=n'_{chip}+1}^{n'_{chip}+n'_{conrol}} \binom{n'_{chip} + n'_{control}}{r} 0.5^{(n'_{chip}+n'_{control})}$$

Where n'_{chip} and $n'_{control}$ are the normalized read counts for the region:

$$n'_{chip} = \frac{n_{chip}}{l_{chip}} \times l_{p_{val}}$$

$$n'_{control} = \frac{n_{control}}{l_{control}} \times l_{p_{val}}$$

Where $l_{p_{val}}$ is the length of the p-value computation window and p refers to the p-value value for the peak. Larger values of $l_{p_{val}}$ increase the significance of regions (See parameter selection). We correct for the p-values using Benjamini-Hochberg procedure to generate the corrected p-values, i.e., q-values:

$$q_i = p_i \times \frac{N_{peaks}}{i}$$

where N_{peaks} is the total number of peak regions and i is the rank of the peak in the peak list sorted with respect to increasing p-value. By default, MUSIC uses q-value cutoff of 0.05. The filtered peaks are reported in BED format with their q-values in the score field.

Multi-Mapability Signal Generation: MUSIC can generate per nucleotide multi-mapability signal profiles. For this it is required to have a read mapping program installed on the system. Currently MUSIC uses bowtie2~\cite{XXX}, a very popular fast read mapping algorithm, by default. MUSIC first fragments all the chromosomes to the read length of interest, maps all the fragments to the genome using bowtie2 with 2 mismatches and reporting of maximum of 5 multimapping positions per fragment. Then MUSIC uses the mapped reads to build the mapability signal profile. The regions with high signal corresponds to regions with low mapability. Then MUSIC processes the mapability profile to store space since it does not require the whole mapability signal profiles. We generated mapability maps for hg19 genome for read lengths of 36, 50, 76, 100, and 200 bps that are available for download with MUSIC.

[[HOW ARE THE THRESHOLDS SELECTED: Add the training F-measure plots.]]

Parameter Selection: There are 3 parameters associated with MUSIC, starting scale window length, ending scale window length, and the p-value computation window length. In order to select the window lengths for broad scale peak calls (H3k36me3, H3k27me3), we used H1HESC human datasets as training dataset (since we used K562 and GM12878 for benchmarking) and ran MUSIC with a large range of parameter sets for the three of F-measure versus percentage overlap between H3k36me3 and H3k27me3 peak calls. This is necessary because we observed that the F-measure increases as the window scales are increased for H3k36me3 dataset. We chose the parameter set that has yields highest F-measure while the overlap percentage is below 1 percent. This parameter set turned out to be l_base=1100 bps, l_end=14000 bps, $l_{p_{val}}$=1750 bps.

Accuracy Measures: For evaluating the accuracy of H3k36me3 peak calls, we computed sensitivity, positive predictive values:

$$Sensitivity = \frac{covg(P \cap G)}{covg(G)}$$

$$PPV = \frac{covg(P \cap G)}{covg(P)}$$

Where $covg(P)$ is the coverage of peaks, $covg(G)$ is the coverage of expressed gene bodies and $covg(P \cap G)$ is the coverage of the overlap between expressed gene bodies and peaks. We combined these two accuracy measures to compute F-measure, computed as:

$$F - measure = \frac{2 \times Sensitivity \times PPV}{(Sensitivity + PPV)}$$

STD2

For H3k4me3 peaks, we used all the promoters (TSS of the transcript ±2500 bps). For these, we use a slightly different approach to compute sensitivity and PPV:

$$Sensitivity = \frac{\#(S \cap P)}{\#(S)}$$

$$PPV = \frac{\#(P \cap S)}{\#(P)}$$

Where $\#(S), \#(P), \#(P \cap S)$ represent number of active promoters, number of peaks, and number of peaks that overlap with active promoters, respectively.

Datasets and Data Processing: We downloaded ENCODE ChIP-Seq from UCSC genome browser. The RNA-seq expression quantifications are downloaded from ENCODE RNA Dashboard. For the transcript quantifications, we used the average RPKM values for the transcripts from two replicates that satisfied the reproducibility criteria that iIDR smaller than 0.1.

SUPPLEMENTARY MATERIAL

Mapability is an important factor for processing genome wide signals. This stems from the fact that the signal levels at region with low mapability will show a systematic decrease at the nucleotide resolution. We used the multi-mapability signal profiles generated by MUSIC (See Methods) and aggregated the signal on different regions (Fig. S1). Promoters and the regions downstream of TSS into the first exon show significantly higher mapability compared to random regions, regions that are upstream into the intergenic side of the genes show significantly lower mapability compared to . In addition, introns show slightly higher mapability compared to random regions and exons show are much more mapable than random regions. Transcription start sites and mid points of exons show almost the same amount of average multi-mapability, 1.2 reads.

Comparison of H3k4me3 ER accuracy with Other Methods:

For H3k4me3, we used the active promoter identification accuracy per top set of peaks of each method for comparison. Although we did not have a negative set for H3k4me3 peaks, unlike H3k36me3, since H3k4me3 is predominantly associated with promoters, we assumed that the top peaks from peak calling will be enriched in active promoters. Starting from the top peaks (sorted with respect to the score reported by each method), we computed the F-measure for promoter identification for each method with changing fraction of coverage of top peaks for the top 30 megabases of the peaks. This way we can evaluate the accuracy of peak calls with changing peak rank. For each peak caller, we sorted the peaks with respect to the reported score. MUSIC tends to perform as one of the best (with MACS) for the accuracy of the top peaks.

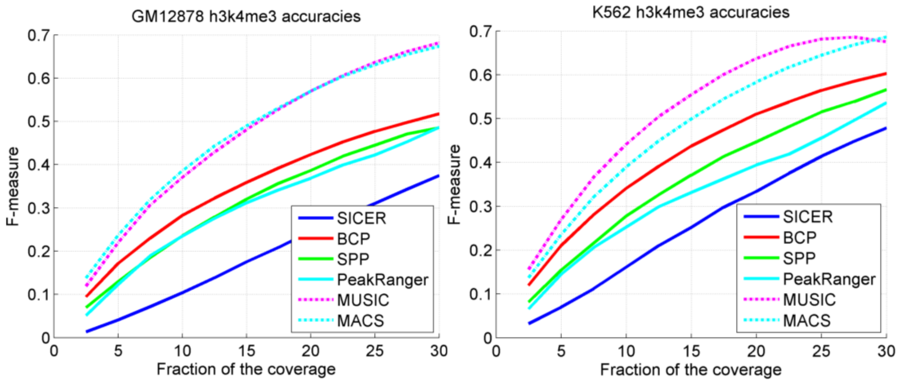
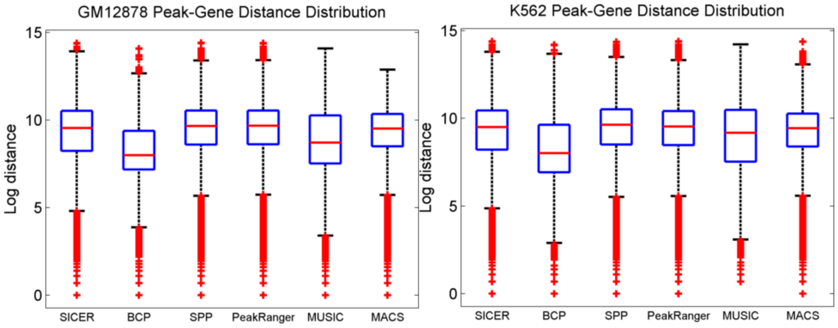


Figure 1: F-measure vs coverage of H3k4me3 peaks in GM12878 (left) and K562 (right)

Next we evaluated the distribution of the distance between the ends of gene bodies and the ends of H3k36me3 peaks to identity whether the peak ends match with the annotated ends of the genes. Figure XXX shows the distribution of smallest peak end to gene end distance for all the peaks for all the methods. The median values are highlighted in the plots to compare the methods with each other. MUSIC has the second smallest median value following BCP.



We also evaluated the reproducibility of the peaks generated by the peak callers. We used the replicates generated by ENCODE with the same HM datasets to assess reproducibility of peak calling. Figure XXX shows the average of fraction of the overlapping regions to the total coverage of each replicate. MUSIC has higher reproducibility for H3k27me3 and H3k36me3 than all other methods except for K562 H3k36me3 dataset, where BCP has slightly higher reproducibility than MUSIC. For K562, MUSIC has highest reproducibility for H3k27me3. For H3k36me3, BCP has slightly higher reproducibility than MUSIC. Overall, MUSIC has higher or comparable reproducibility with respect to other peak callers.